PB86-188182

Clustering of Rare Events

North Carolina Univ. at Chapel Hill

Prepared for

Health Effects Research Lab.
Research Triangle Park, NC

1983

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA (Please read Instructions on the reverse before completi)		
1. REPORT NO. EPA/600/J-83/338	2.	3. PB86-188182
4. TITLE AND SUBTITLE CLUSTERING OF RARE EVENTS (Journal Version)		5. REPORT DATE 1983
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) [1]M.J. Symons, R.C. Grimson and Y.C. Yuan		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS [1]Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC 27514		10. PROGRAM ELEMENT NO.
		11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory Office of Research and Development U.S. Environmental Protection Agency Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED Journal Article
		14. SPONSORING AGENCY CODE EPA - 600/11
15. SUPPLEMENTARY NOTES Biometrics 39, 193-205 March 1983		

16. ABSTRACT

The clustering of cases of a rare disease is considered. The number of events observed for each unit is assumed to have a Poisson distribution, the mean of which depends upon the population size and the cluster membership of that unit. Here a cluster consists of those units that are homogeneous in their rate of occurrence of the rare events under study. A sample of units is modeled by a mixture of Poisson distributions, one for each cluster, the mixing parameters being the proportions of all units represented by the components of the mixture. Maximum likelihood and Bayes approaches are employed to determine criteria for separating a sample into groups of units with homogeneous rates. A likelihood ratio test for the significance of a two-component mixture is presented as an example. The performance of the criteria is illustrated with data on the spatial occurrence of sudden infant deaths (SIDs) in North Carolina counties over a four-year period. The results suggest that the practice of dividing the counties into high- and low-risk categories on the basis of the ordered rates alone should be questioned. Tests based upon combinatorial methods are also presented to examine the significance of the number of contiguous counties among those with high rates.

17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	14
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

BIOMETRICS 39, 193–205
March 1983

EPA/600/J-83/338
JOURNAL ARTICLE

Clustering of Rare Events

M. J. Symons, R. C. Grimson and Y. C. Yuan

Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27514, U.S.A.

SUMMARY

The clustering of cases of a rare disease is considered. The number of events observed for each unit is assumed to have a Poisson distribution, the mean of which depends upon the population size and the cluster membership of that unit. Here a cluster consists of those units that are homogeneous in their rate of occurrence of the rare events under study. A sample of units is modeled by a mixture of Poisson distributions, one for each cluster, the mixing parameters being the proportions of all units represented by the components of the mixture. Maximum likelihood and Bayes approaches are employed to determine criteria for separating a sample into groups of units with homogeneous rates. A likelihood ratio test for the significance of a two-component mixture is presented as an example. The performance of the criteria is illustrated with data on the spatial occurrence of sudden infant deaths (SIDs) in North Carolina counties over a four-year period. The results suggest that the practice of dividing the counties into high- and low-risk categories on the basis of the ordered rates alone should be questioned. Tests based upon combinatorial methods are also presented to examine the significance of the number of contiguous counties among those with high rates.

1. Introduction

Clustering of a disease in time, space or both should offer clues to the etiology of the disease. However, in the absence of etiologic knowledge, the choice of an appropriate statistical structure for the clustering is obscure. Also, the power of different clustering models depends on the potentially unknown mode of disease transmission, or on the combination of risk factors for the disease. Consequently, the detection of clustering in time, space or time–space units, perhaps by a statistical test, and the labeling of the cluster membership of each unit by statistical modeling, are two important related aspects of the investigation of clustering. Previous effort has been largely spent on the development of tests for clustering. A review of these tests, and of statistical criteria useful in identifying the cluster membership of each unit, will now be presented.

Ederer, Myers and Mantel (1964) and Mantel, Kryscio and Myers (1976) developed a test for temporal clusters of disease within geographic units, and showed how to increase the power of the test by aggregating the units. Wallenstein (1980) reviewed and extended a similar test for disease clustering. Additional work on statistical tests and models for spatial clustering of either cases of disease or geographic units with high incidence of disease were discussed and referenced by Batty and Sammons (1978) and Haggett, Cliff and Frey (1977).

Knox (1963, 1964) studied the clustering of disease in both space and time, and dealt specifically with space–time interactions. Many others have also examined this problem, for example, P. K. Roberson (in a Ph.D. dissertation at the University of Washington, 1979), Mantel (1967) and Klauber (1975). Basically, space–time interaction is manifest when more cases are observed than would be expected to occur by chance alone within a short period of time *and* the cases are in close proximity to one another. Increasing or decreasing trends over

Key words: Poisson mixture; Bayesian estimation; Likelihood ratio; Combinatorics; Sudden infant death.

NTIS is authorized to reproduce and sell this report. Permission for further reproduction must be obtained from the copyright owner.

time in some of the counties within a country or state would be examples of space–time interaction.

Most methodology in the literature has focused on tests for clustering. Little attention has been given to the statistical modeling of disease incidence, other than to applications of stochastic process to the propagation of diseases; see Grimson (1979). Modeling may reveal informative epidemiologic characteristics or patterns of disease processes. It may be argued that most disease incidence does not behave 'purely at random' in time or space, although such incidence may often appear random when only observational data are available. This paper focuses on the identification and characterization of disease clusters, as well as on tests for the existence of clustering.

Some disease-cluster tests are based only on the magnitude of the rates of disease occurrence, their stability being ignored, and a group of high-risk units is established by a visually identified gap in the ordered rates. Unless the rates are known to have the same precision, such an apparent gap may be misleading as a basis for classification. An important feature of the criteria proposed in this paper is the use, in the classification process, of the precision of the rate for each unit, as well as its magnitude. The larger the population base for the rate, of course, the more stable the rate will be, and also the more reliable will be the classification of the corresponding unit.

As an example, consider the sudden infant death syndrome (SIDS) and the problem of identifying the counties at high risk to SIDS in North Carolina. The data available are the number of sudden infant deaths (SIDs) and the number of live births for each of four years, but the total for the four years is taken as descriptive of the disease clustering. Geographic proximity of the identified high-risk counties is then considered. The similarities in geographic, economic, demographic, or sociologic characteristics of those counties in North Carolina with larger rates of SIDs may reveal a risk factor for this mysterious killer. Alternatively, examination of cases in the high-rate counties may suggest physical transmission routes for organisms or viral agents, or familial commonalities of genetic origin. The identification of the high-risk counties is a preliminary step that expedites subsequent investigation and analysis. For this purpose a mixture of Poisson distributions is considered as a clustering model.

Before proceeding, a few comments on terminology seem appropriate. In the SIDS example a 'cluster' denotes a group of counties with similar rates. The process of labeling these counties is probably more accurately described as 'simultaneous classification', following Geisser (1966), than as 'clustering.' Geographic proximity or contiguity of the high-risk counties also seems to be implied by the natural meaning of 'cluster.' 'Regionalization', as used by Haggett *et al.* (1977), is probably a more precise description of the joint occurrence of a high SIDS rate and contiguous county boundaries. Although more exact terms are available, 'cluster' and 'clustering' are conveniently used in a variety of contexts and so will be employed here.

2. Statistical Model and Consequent Criteria

We shall use the SIDS example for orientation. The counties are presumed to be a mixture of two groups, 'normal'- and 'high'-risk, the generalization to more groups being immediate. Since the events of interest are taken to be rare, the number of events, d_i, in the ith county, which has N_i individuals at risk over time t_i, is assumed to have a Poisson distribution. As the product $N_i t_i$, denoted by V_i, represents the volume or mass of the population at risk, the distribution of the number of events is given by

$$P(d_i \mid V_i, \lambda_g) = (V_i\lambda_g)^{d_i}\exp(-V_i\lambda_g)/d_i!, \tag{1}$$

where $i = 1, \ldots, n$ indexes the counties and λ_g is the population rate in the gth group, $g = 1$ (normal risk) or $g = 2$ (high risk). The distribution of the number of SIDs is conditional on whether the county has a normal or high risk and on the exposed population mass of the

county. A randomly selected county is normal-risk with probability π_1 and high-risk with probability $\pi_2 = 1 - \pi_1$. The unconditional probability of d_i is then

$$f(d_i \mid \pi_g, \lambda_g, V_i) = \sum_{g=1}^{G} \pi_g P(d_i \mid V_i, \lambda_g), \tag{2}$$

where $G = 2$ in the SIDS example, but may be larger in other applications.

The cluster problem is formulated as one of estimating the unknown mixture-component origin of each d_i, denoted by $z_i = g$, or equivalently asserting that d_i comes from the gth component of the mixture. The *a priori* probability that $z_i = g$ is π_g. The likelihood of the data is required in order to estimate the n components of the vector $\mathbf{z} = (z_1, \ldots, z_n)$ by maximum likelihood or by a Bayesian approach. Let the vector $\mathbf{d}$ denote the n observations d_i, $\mathbf{V}$ be the vector of corresponding at-risk volumes V_i, and $\boldsymbol{\theta}$ be the vector of parameters $(\pi_1, \ldots, \pi_G, \lambda_1, \ldots, \lambda_G)$. The likelihood of the data for the n counties, with specified allocation $\mathbf{z}$, parameters $\boldsymbol{\theta}$ and masses $\mathbf{V}$, is

$$L(\mathbf{d} \mid \boldsymbol{\theta}, \mathbf{z}, \mathbf{V}) = \left\{\prod_{g=1}^{G} (\pi_g^{n_g})\right\} \exp\left\{-\sum_{g} \sum_{C_g} V_i \lambda_g + \sum_{g} \sum_{C_g} d_i \ln(V_i \lambda_g)\right\}. \tag{3}$$

The n_g denotes the number of observations allocated to the gth component by $\mathbf{z}$, and C_g is the collection of these observations. The first factor on the right-hand side (3) represents the likelihood of the *number* of observations assigned to each group, and the remainder represents the likelihood of *which* of these numbers of observations are assigned to each group.

Following a maximum likelihood (ML) approach, we obtain $\hat{\mathbf{z}}$, the maximum likelihood estimate (MLE) of $\mathbf{z}$, as the allocation that maximizes (3). This is accomplished by replacing the parameters $\boldsymbol{\theta}$ by their MLEs, given an allocation of the n quantities d_i to the G components, and then seeking the maximum of (3) over the G^n possible allocations. (There are obvious practical difficulties in this search, and the computational details are considered with the example in §3.) Given an allocation $\mathbf{z}$, the MLEs of the parameters are the natural ones, namely

$$\hat{\pi}_g = n_g / n \tag{4}$$

and

$$\hat{\lambda}_g = \sum_{C_g} d_i / \sum_{C_g} V_i. \tag{5}$$

The allocation $\hat{\mathbf{z}}$ that maximizes (3) is equivalent to the partition of the n observations into G groups that minimizes the negative of the natural logarithm of (3), specifically,

$$\sum_{i=1}^{n} d_i - \sum_{g=1}^{G} \sum_{C_g} d_i \ln(V_i \hat{\lambda}_g) - \sum_{g=1}^{G} n_g \ln n_g. \tag{6}$$

A Bayesian approach requires the specification of a prior distribution, $p(\boldsymbol{\theta})$, for $\boldsymbol{\theta}$ and the averaging over the uncertainty in the unknown parameters, since the allocation vector $\mathbf{z}$ is of primary interest. Generally, the integrated likelihood is

$$L(\mathbf{d} \mid \mathbf{z}, \mathbf{V}) = \int L(\mathbf{d} \mid \boldsymbol{\theta}, \mathbf{z}, \mathbf{V}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{7}$$

where the integration is over the parameter space of $\boldsymbol{\theta}$. The mode of (7) is taken as the Bayes estimate, $\tilde{\mathbf{z}}$, of the optimal allocation. In accordance with the discussion by Lindley (1965) and Jeffreys (1961), the following vague prior is chosen to delineate the parameter space:

$$p(\boldsymbol{\theta}) = p(\pi_1, \ldots, \pi_G, \lambda_1, \ldots, \lambda_G) \propto \left(\prod_{g=1}^{G} \pi_g^{-1}\right)\left(\prod_{g=1}^{G} \lambda_g^{-1}\right). \tag{8}$$

The likelihood (3) is averaged over the prior distribution (8), as in (7). This procedure involves the normalization constants from a gamma distribution and a Dirichlet distribution. For more details, see Johnson and Kotz (1969), Binder (1978) and Symons (1981). The resulting Bayesian optimal allocation, $\tilde{\mathbf{z}}$, is equivalent to that partition of the data into G groups which minimizes the criterion

$$\sum_{g=1}^{G}\left(\sum_{C_g} d_i\right)\ln\left(\sum_{C_g} V_i\right) - \sum_{g=1}^{G}\ln\left\{\Gamma\left(\sum_{C_g} d_i\right)\right\} - \sum_{g=1}^{G}\ln\{\Gamma(n_g)\}\,. \tag{9}$$

A mixture of multivariate normals, approached in this way, yields criteria that are modifications of existing clustering criteria; see Symons (1981). No analogous recognition is apparent in the present application, although a sensitivity to disparate-sized clusters can be expected with these criteria, due to the sums over g in (6) and (9) that involve the cluster sizes, n_g.

The criteria, (6) from ML estimation and (9) from Bayes theory, are derived from the likelihood (3). Terminology and notation for this likelihood should be discussed briefly. 'Joint likelihood' and $L(\mathbf{d}, \mathbf{z} \mid \boldsymbol{\theta}, \mathbf{V})$ recognize both $\mathbf{d}$ (observable) and $\mathbf{z}$ (unobservable) as data and coincide with the description following (3). However, 'conditional likelihood' and $L(\mathbf{d} \mid \boldsymbol{\theta}, \mathbf{z}, \mathbf{V})$ emphasize that evaluation of the likelihood is possible only at a specified value, rather than at an observed value, of the allocation vector $\mathbf{z}$. A 'marginal' or 'unconditional' likelihood can then be written as

$$L(\mathbf{d} \mid \boldsymbol{\theta}, \mathbf{V}) = \prod_{i=1}^{n}\left\{\sum_{g=1}^{G}\pi_g\,(V_i\lambda_g)^{d_i}\exp(-V_i\lambda_g)/d_i!\right\}, \tag{10}$$

and is the basis of another familiar classification approach. The likelihood (10) can be obtained by summing (3) over all G^n possible values of $\mathbf{z}$.

A likelihood ratio approach, analogous to that of discriminant analysis in Anderson (1958, Chapter 6) with an implicit assumption of equal costs of misclassification, is considered next. If, for the likelihood (10), $\boldsymbol{\theta}$ is known or MLEs of the parameters with large samples are available, the n counties can be classified independently from the Poisson component of the mixture (2) with the largest scaled probability, scaling being by the proportion of counties represented by the component. In particular, an alternative large-sample likelihood ratio procedure is to classify d_i to the mixture component g when

$$\hat{\pi}_g P(d_i \mid V_i, \hat{\lambda}_g) = \max_{1\le j\le G}\{\hat{\pi}_j P(d_i \mid V_i, \hat{\lambda}_j)\}, \tag{11}$$

where $\hat{\lambda}_j$ and $\hat{\pi}_j$ are the large-sample MLEs or known parameter values.

The required ML parameter estimates for (10) with the likelihood ratio procedure (11) can be obtained by an iterative scheme, for example, that described in the technical report by E. B. Kaplan and R. C. Elston (Institute of Statistics Mimeo Series No. 823, University of North Carolina, Chapel Hill, 1972). Initial estimates for such procedures may be determined by preliminary descriptive analysis of the data or, more precisely, by the method of moments; see Rider (1961) and Cohen (1963). Johnson and Kotz (1969) have reviewed these approaches. Alternatively, initial estimates based upon $\hat{\mathbf{z}}$ or $\tilde{\mathbf{z}}$ from Criterion (6) or (9), respectively, may be employed. These are biased parameter estimates, as Marriott (1975) noted. However, in a technical report (Institute of Statistics Mimeo Series No. 880, University of North Carolina, Chapel Hill, 1973), M. J. Symons found that as initial estimates for an iterative ML scheme they seem to work quite well for a mixture of multivariate normals, probably because the bias in the estimates of the means tends to exaggerate slightly the separation of the mixture components.

A general guide to the use of these criteria can be based, in part, upon the available sample

size. The Bayes estimates of the parameters in the likelihood (10) with vague prior information will converge to the MLEs with large samples; see Lindley (1965) or Cox and Hinkley (1975). Given the desirable large-sample properties of MLEs, the general optimality of the likelihood ratio criterion, and the computation difficulties in optimizing criterion (6) or (9) over G^n allocations, the likelihood ratio procedure in (11) is preferable for large samples. The required sample size for this preference is difficult to specify, since the amount of separation in the rates and the number and relative size of the component groups affects the precision with which the parameters θ can be estimated. With smaller sample sizes, either Criterion (6) or (9) may be more reliable, as was indicated for mixtures of multivariate normals in the 1973 technical report by Symons. A comparison of the results of using these criteria in a specific example, is presented in §3.

3. Sudden Infant Death Syndrome: An Example

Sudden infant death syndrome has become a leading category of postneonatal death in the last decade. Except for a seasonal pattern of excessive mortality in the winter months, shown by Goldberg and Stein (1978), clustering of SIDs in calendar time (years) or space remains undocumented, according to Atkinson (1978). The seasonal excess of SIDs in the winter months would be consistent with a familial genetic factor, such as a defect in the respiratory system which could be fatal in some circumstances for infants, but may not be serious for adults. The review of the SIDS literature by Atkinson (1978) includes an attempt to cluster SIDs by using the Ederer–Myers–Mantel method. No temporal clustering within county by single years of calendar time was found to be significant. Atkinson used the number and rate of SIDs by county and single years over the period 1 July 1974 through 30 June 1978. The number of live births by county for the same calendar period was added to Atkinson's data, with the kind cooperation of the North Carolina Department of Human Resources. The results are shown in Table 1.

The results of clustering the 100 counties of North Carolina into high- and normal-SIDS-risk counties by the criteria of §2 are given in Table 2. The 20 counties with the largest SIDS death rates per 1000 live births are listed in decreasing order of that rate. The other 80 counties were classified as normal-risk ('Lo' in Table 2) by all the criteria. The criteria employed were the conditional maximum likelihood (6), the conditional Bayes (9), and the unconditional likelihood ratio (11). As can be seen from Table 2, the classification can vary considerably according to the criterion selected and the local optimum for Criteria (6) and (9). However, these classifications of the counties are not affected by removing the county with the highest rate, Anson County. Also, Criteria (6) and (9) provide identical classifications for these data.

An important feature of these criteria is that the classification of a county as high-risk is based not only on the SIDS rate, but also on the number of live births. As the variance of the estimate of the rate in a Poisson process decreases with increasing volume of the population at risk, here given by the number of live births in the four-year period, it is reassuring to note that a county with a high rate based upon few births is not necessarily classified as high-risk. For example, Swain county had a rate of 4.4 SIDs per 1000 live births, but based upon only three SIDs and 675 live births from 1974 to 1978. This county was classified as normal-risk by all the criteria. A more informal rule, based, for example, upon a gap in the ordered SIDS death rates for the counties, would have classified Swain county as high-risk if more than 10% of the counties were thought to be high-risk.

Unfortunately, the results for these criteria also serve as a counter-example to a natural suggestion for finding the optimum of Criterion (6) or (9) by splitting the ordered SIDS death rates. Instead of only $n - 1$ partitions being examined for a split into two non-empty groups, $2^n - 2$ allocations should be examined to assure the global optimization of Criterion (6) or

Table 1

Number of live births, and number and rate of infant deaths with mention of sudden infant death syndrome† by county, North Carolina, 1974–1978‡*

County	No. of live births	No. of SIDs	Rate*	County	No. of live births	No. of SIDs	Rate*	County	No. of live births	No. of SIDs	Rate*
Alamance	4 672	13	2.8	Franklin	1 399	2	1.4	Pamlico	542	1	1.8
Alexander	1 333	0	0.0	Gaston	9 014	11	1.2	Pasquotank	1 638	3	1.8
Alleghany	487	0	0.0	Gates	420	0	0.0	Pender	1 228	4	3.3
Anson	1 570	15	9.6	Graham	415	0	0.0	Perquimans	484	1	2.1
Ashe	1 091	1	0.9	Granville	1 671	4	2.4	Person	1 556	4	2.6
Avery	781	0	0.0	Greene	870	4	4.6	Pitt	5 094	14	2.7
Beaufort	2 692	7	2.6	Guilford	16 184	23	1.4	Polk	533	1	1.9
Bertie	1 324	6	4.5	Halifax	3 608	18	5.0	Randolph	4 456	7	1.6
Bladen	1 782	8	4.5	Harnett	3 776	6	1.6	Richmond	2 756	4	1.5
Brunswick	2 181	5	2.3	Haywood	2 110	2	0.9	Robeson	7 889	31	3.9
Buncombe	7 515	9	1.2	Henderson	2 574	5	1.9	Rockingham	4 449	16	3.6
Burke	3 573	5	1.4	Hertford	1 452	7	4.8	Rowan	4 606	3	0.7
Cabarrus	4 099	3	0.7	Hoke	1 494	7	4.7	Rutherford	2 992	12	4.0
Caldwell	3 609	6	1.7	Hyde	338	0	0.0	Sampson	3 025	4	1.3
Camden	286	0	0.0	Iredell	4 139	4	1.0	Scotland	2 255	8	3.5
Carteret	2 414	5	2.1	Jackson	1 143	2	1.8	Stanly	2 356	5	2.1
Caswell	1 035	2	1.9	Johnston	3 999	6	1.5	Stokes	1 612	1	0.6
Catawba	5 754	5	0.9	Jones	578	1	1.7	Surry	3 188	5	1.6
Chatham	1 646	2	1.2	Lee	2 252	5	2.2	Swain	675	3	4.4
Cherokee	1 027	2	1.9	Lenoir	3 589	10	2.9	Transylvania	1 173	3	2.6
Chowan	751	1	1.3	Lincoln	2 216	8	3.6	Tyrrell	248	0	0.0
Clay	284	0	0.0	McDowell	1 946	5	2.6	Union	3 915	4	1.0
Cleveland	4 866	10	2.1	Macon	797	0	0.0	Vance	2 180	4	1.8
Columbus	3 350	15	4.5	Madison	765	2	2.6	Wake	14 484	16	1.1
Craven	5 868	13	2.2	Martin	1 549	2	1.3	Warren	968	4	4.1
Cumberland	20 366	38	1.9	Mecklenburg	21 588	44	2.0	Washington	990	5	5.1
Currituck	508	1	2.0	Mitchell	671	0	0.0	Watauga	1 323	1	0.8
Dare	521	0	0.0	Montgomery	1 258	3	2.4	Wayne	6 638	18	2.7
Davidson	5 509	8	1.5	Moore	2 648	5	1.9	Wilkes	3 146	4	1.3
Davie	1 207	1	0.8	Nash	4 021	8	2.0	Wilson	3 702	11	3.0
Duplin	2 483	4	1.6	New Hanover	5 526	12	2.2	Yadkin	1 269	1	0.8
Durham	7 970	16	2.0	Northampton	1 421	9	6.3	Yancey	770	0	0.0
Edgecombe	3 657	10	2.7	Onslow	11 158	29	2.6				
Forsyth	11 858	10	0.8	Orange	3 164	4	1.3	Total for North Carolina	329 962	667	2.0

* Deaths per 1000 live births.

† Infant deaths between 28 days and one year of age, in which the underlying cause of death is coded as 7950 or in which sudden infant death syndrome was mentioned on the death certificate although the underlying cause was not coded as such.

‡ Defined as 1 July 1974–30 June 1978.

Table 2

Results with three criteria for the 20 North Carolina counties with the largest SIDS death rate, 1974–1978*

County	No. of live births	No. of SIDs	Rate per 1000	Rank	Classification of counties by criterion				
					Conditional ML, (6)†		Conditional Bayes, (9)‡		Likelihood ratio (11)
					1st	2nd	1st	2nd	
Anson	1570	15	9.6	1	Hi	Hi	Hi	Hi	Hi
Northampton	1421	9	6.3	2	Hi	Hi	Hi	Hi	Hi
Washington	990	5	5.1	3	Lo	Lo	Lo	Lo	Hi
Halifax	3608	18	5.0	4	Hi	Hi	Hi	Hi	Hi
Hertford	1452	7	4.8	5	Lo	Hi	Lo	Hi	Hi
Hoke	1494	7	4.7	6	Lo	Hi	Lo	Hi	Hi
Greene	870	4	4.6	7	Lo	Lo	Lo	Lo	Hi
Bertie	1324	6	4.53	8	Lo	Lo	Lo	Lo	Hi
Bladen	1782	8	4.49	9	Lo	Hi	Lo	Hi	Hi
Columbus	3350	15	4.48	10	Hi	Hi	Hi	Hi	Hi
Swain	675	3	4.4	11	Lo	Lo	Lo	Lo	Lo
Warren	968	4	4.1	12	Lo	Lo	Lo	Lo	Lo
Rutherford	2992	12	4.0	13	Lo	Hi	Lo	Hi	Hi
Robeson	7889	31	3.9	14	Hi	Hi	Hi	Hi	Hi
Lincoln	2216	8	3.61	15	Lo	Lo	Lo	Lo	Hi
Rockingham	4449	16	3.60	16	Lo	Hi	Lo	Hi	Hi
Scotland	2255	8	3.5	17	Lo	Lo	Lo	Lo	Hi
Pender	1228	4	3.3	18	Lo	Lo	Lo	Lo	Lo
Wilson	3702	11	3.0	19	Lo	Lo	Lo	Lo	Lo
Lenoir	3589	10	2.9	20	Lo	Lo	Lo	Lo	Lo

* The other 80 of the 100 North Carolina counties were classified as "Lo" risk by all three criteria.

† Criterion value for the smallest (second smallest) ML criterion (6) was equal to −1319.92 (−1319.58).

‡ Criterion value for the smallest (second smallest) Bayes criterion (9) was equal to 4440.40 (4441.08).

(9). Of course, this is impractical for even moderate-sized samples, and partial searches need to be employed. The program due to McRae (1971) was adapted for this purpose, but it does not guarantee a global minimum.

An alternative procedure is to classify a county as high-risk if the *P*-value for the observed number of SIDs is less than some prespecified level. With a single Poisson distribution as the null hypothesis, the *P*-value is computed for each county as the sum of Poisson probabilities corresponding to values at least as large as the observed number of SIDs. The 1974–1978 experience for all of North Carolina, namely 2.0 SIDs per 1000 live births, is used to determine the expected value, $V_i/500$, of the null-hypothesis Poisson distribution for the ith county. The classification of counties by this informal outlier-screening device is the same as that given by the smallest values of Criteria (6) and (9) when a tail probability of .01 is prespecified. If a tail probability of .05 is prespecified, the outlier classification rule agrees with the second smallest values of these two criteria. However, as a general rule such a correspondence would not be expected.

4. Statistical Significance

4.1 *Two-Component Mixture or Single Poisson Distribution*

Once the number of clusters is specified, the criteria in §2 can be used to assign each observation to a group. However, to decide upon the number of components in a mixture is

a very difficult problem. The reason for this difficulty is that the various choices for the number of clusters, G, the weights, π_g, and the rates, λ_g, provide tremendous flexibility in modeling a data set.

The likelihood ratio test of a null hypothesis $H_0: G = G_0$, versus an alternative hypothesis $H_1: G = G_1$, where G_0 and G_1 are specified integer values, is a natural choice as a test criterion. For example, with $G_0 = 1$ and $G_1 = 2$, the likelihood ratio (LR) is

$$\mathrm{LR} = \prod_{i=1}^{n} \left\{ \frac{(V_i\hat{\lambda}_0)^{d_i}\exp(-V_i\hat{\lambda}_0)}{\hat{\pi}_1(V_i\hat{\lambda}_1)^{d_i}\exp(-V_i\hat{\lambda}_1) + \hat{\pi}_2(V_i\hat{\lambda}_2)^{d_i}\exp(-V_i\hat{\lambda}_2)} \right\}, \tag{12}$$

where $\hat{\lambda}_0$, $(\hat{\lambda}_1, \hat{\lambda}_2)$ and $\hat{\pi}_1(\hat{\pi}_2 = 1 - \hat{\pi}_1)$ are the ML parameter estimates for the null Poisson distribution and the hypothesized mixture of Poisson distributions, respectively. No summary sufficient statistics of dimension less than n exist for the parameters of the mixture model, and analytical work is necessarily complex. In addition, the hypothesis of one versus two groups can be restated as π_1 being zero or unity with $\pi_2 = 1 - \pi_1$, or as $\lambda_1 = \lambda_2$ for any value of π_1. In view of these equivalencies on the boundary of the parameter space of π_1, the usual result of Wilks (1938) that minus twice the log likelihood ratio follows a chi square distribution may not apply.

In an attempt to assess the likelihood ratio test statistic given by (12) for the observed data in Table 1, a simulation was performed, on the assumption of the null hypothesis that a single rate is common to all the North Carolina counties. The IMSL generator GGPOS by Snow (1968) was used to simulate the Poisson values. The grid-search routine of Kaplan and Elston's 1972 Technical Report was used to maximize the denominator of (12). In each of the 2000 samples generated, the expected number of SIDs for the ith county was $V_i/500$, corresponding to a rate of 2.0 SIDs per 1000 live births. The observed value of minus twice the log-likelihood ratio was 61.15, which exceeded the largest of the 2000 simulated values, specifically 16.269, suggesting a P-value of much less than .0005.

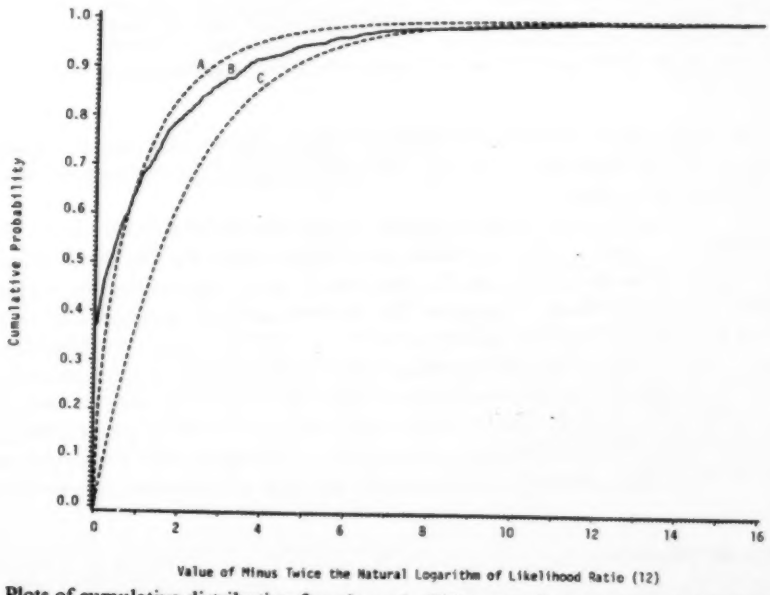

Figure 1. Plots of cumulative distribution functions: A, Chi square distribution with 1 df; B, Simulated distribution of minus twice the log-likelihood ratio from (12); C, Chi square distribution with 2 df.

In Fig. 1 the simulated cumulative distribution of -2 ln (LR), with LR given by (12), is plotted along with the corresponding cumulative chi square distributions with 1 df and 2 df. As a preliminary conclusion and for the purpose of this example, it seems reasonable to approximate the upper tail of the distribution of -2 ln (LR) by a chi square distribution with 2 df. Therefore, the statistical significance of the observed value of -2 ln (LR) = 61.15 is not in question. However, the distribution of this statistic under the null hypothesis is not so clear. The fit is inferior in the lower tail, raising the question of whether the null distribution of -2 ln (LR) could be approximated by a chi square distribution with fractional degrees of freedom. Much of the lack of fit is due to the large number of zeros observed for -2 ln (LR) in the simulation. Since the denominator of (12) contains the numerator as a special case of its parameter values, this likelihood ratio should always be less than unity. The point is that the ratio can equal unity, so a zero value of -2 ln (LR) can arise. The simulated samples were from a Poisson model, in which the mean equals the variance. However, in the model which involves a mixture of Poisson distributions, the variance is greater than the mean. For simulated samples where the sample mean exceeds the sample variance, the best-fitting mixture may well be a single Poisson, i.e. on the boundary of the parameter space.

The simulation results for North Carolina counties suggest that a mixture with two Poisson components, rather than a single Poisson distribution, is justifiable. Thus it is reasonable to label the counties high- or normal-risk, as in Table 2. These results are displayed in Fig. 2. The high-risk counties, determined by the smallest value of (6) and (9), are indicated by the darkest shading. Those labeled as high-risk on the basis of the second smallest values of Criteria (6) and (9) are indicated by the medium shading plus the darkest shading. The lightest plus the two darker shadings represent the counties labeled as high-risk by the likelihood ratio criterion (11). Those counties marked with dots are those remaining of the 20 counties with the largest SIDS rates. Note that Swain county is the westernmost county identified by dots in Fig. 2; it is classified as normal-risk by these three criteria as its rate is based on relatively few live births. The rate for Swain county is the 11th largest of the 100 counties.

The MLEs and standard errors of the parameters of the two-component Poisson mixture in the denominator of the likelihood ratio (12) are $\hat{\pi}_1 = 0.7969(\pm 0.1298)$, $\hat{\lambda}_1 = 0.00169(\pm 0.00015)$ and $\hat{\lambda}_2 = 0.00380(\pm 0.00087)$. Recall that π_1 is the proportion of counties at normal risk, so approximately 20% of the counties were estimated as high-risk. However, only 15 were indicated by Criterion (11) as high-risk. The rate among the high-risk counties was estimated to be 3.80 ± 0.87 SIDs per 1000 live births, while the rate for normal-risk counties was estimated at 1.69 ± 0.15 SIDs per 1000 live births.

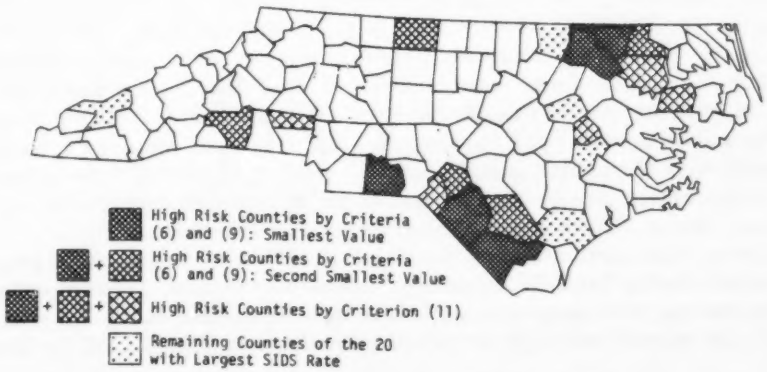
High Risk Counties by Criteria (6) and (9): Smallest Value
High Risk Counties by Criteria (6) and (9): Second Smallest Value
High Risk Counties by Criterion (11)
Remaining Counties of the 20 with Largest SIDS Rate

Figure 2. Map of North Carolina showing the results from Table 2.

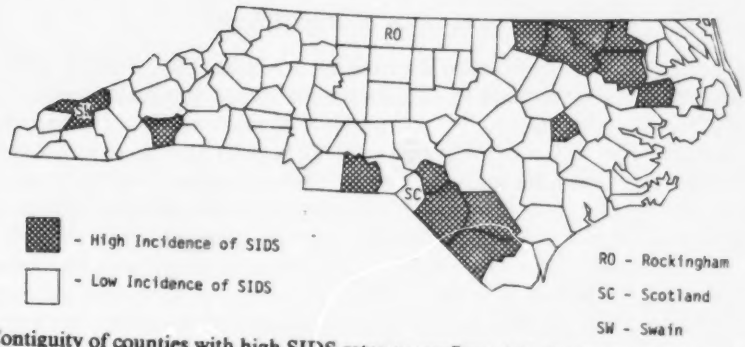

Figure 3. Contiguity of counties with high SIDS rates as confirmed by the Monte Carlo regionalization test.

4.2 *Combinatorial Tests for Geographical Contiguity*

Given that it is reasonable to assign labels of high- and low-risk to the counties of North Carolina, a subsequent question is whether or not the high-risk counties are randomly distributed across the state. Combinatorial arguments have been developed to enable deviations from a random distribution in space to be examined. Grimson, Wang and Johnson (1981) have shown that North Carolina counties with the highest SIDS rates tend to form a contiguous region in the northeastern area of the State and a second, smaller region of a slightly lower, yet still high, rate in the southern part of the State; see Fig. 3. The statistical significance of this geographical clustering of counties was based upon 24 000 simulations generated by randomly sampling m counties from among the 100 for each value of m from two to 25, producing the distribution of the number of adjacencies. Consequently, if the number of high-risk counties is specified, the P-value corresponding to any number of adjacencies is available from Table 1 of Grimson *et al.* (1981).

Haggett *et al.* (1977, Vol. 2, Chapter 11) reviewed several tests for this type of spatial clustering, their tests being based upon graph-theoretic principles and distributional assumptions rather than simulation. Ohno, Aoki and Aoki (1979) also presented a Monte Carlo test of the same type as Grimson *et al.* (1981) and suggested large-sample approximations.

According to the method of Grimson *et al.* (1981), the counties are ranked from the highest to the lowest SIDS rate, as in Table 2. Then one proceeds sequentially down the list of ranked highest-risk counties and notes the P-value of the number of observed adjacencies, corresponding to $m = 2, 3, 4, \ldots, 25$. A low P-value for a specified number of counties with a high SIDS rate is regarded as an indication of geographic clustering, or regionalization, among those counties. For $m = 14$, corresponding to Robeson County, the P-value is .00005 and there is a gap in the SIDS rate from this county to that of the 15th county, Lincoln. The map in Fig. 3 shows the geographic distribution of these 14 counties, which corresponds well with the results from the likelihood ratio criterion shown in Fig. 2. Note the differences due to the stability of the rates between the classification by this criterion in Fig. 2 and the 14 counties with the largest SIDS rates in Fig. 3. Not only may high rates based upon few SIDs and live births be classified as normal-risk, as for Swain county, but moderate SIDS rates based upon many live births may be classified as high-risk, as for Rockingham and Scotland counties. The locations of Swain, Rockingham and Scotland counties are shown in Fig. 3.

The clustering, in the sense of the counties with high SIDS rates forming local geographic pockets, is summarized in Table 3 using the regionalization criterion of Grimson *et al.* (1981). The visual impression of the geographic pockets of high-risk counties in Fig. 3 is immediate, even though some counties with high rates based upon relatively few SIDs and live births are

Table 3
Evidence that counties classified as high-risk for SIDS by selected criteria are contiguous

Criteria for classifying counties as high-risk	Number of counties classified	Number of county adjacencies	*P*-value from test for contiguity (Grimson *et al.*, 1981)
Criteria (6) and (9), smallest value	4	2	.030
Criteria (6) and (9), second smallest value	10	6	.015
Criteria (11)	15	12	.003

included. The small *P*-values in Table 3, corresponding to the groups of high-risk counties identified by the criteria in Table 2, support this impression.

If the analysis of Grimson *et al.* (1981) is repeated, with the counties having unstable rates being regarded as neither high- nor normal-risk, there is still evidence of regionalization in these data. A clustering of counties with high SIDS risk then seems to be a reasonable conclusion. Therefore, it seems justified to use the criteria put forward in this paper to specify those counties of high SIDS risk and then to look for a tendency toward regionalization. Other analyses which seek correlations for the high-risk label of each county, or case–control examination of SIDs in these high-risk counties, are natural subsequent steps.

5. Discussion

The three criteria developed in §2, namely, (6), (9) and (11), and other clustering strategies that may be simpler to compute, are probably best *not* viewed as competitors. Rather, a reasonable view is that no single clustering criterion is likely to produce reliable results in most situations. The consistency and stability of clusters produced by different approaches is a safeguard against potentially spurious results from a single approach. Everitt (1979) made essentially this same point; it seems worthy of repetition.

Although the conditions of Wilks (1938) for the large-sample distribution of the likelihood ratio in (12) are not met in this application, it is noteworthy that the *upper* percentiles of the empirical distribution are approximated by those of a chi square distribution with 2 df—the difference between the dimension of the mixture-model parameter space (three) and that of a single Poisson (one). Counting the degrees of freedom as the number of independent constraints placed upon the mixture-model parameter space is not so clearly done in this situation. The large-sample distribution of the likelihood ratio distribution for mixtures needs further study. However, on the basis of this simulation and the discussion by Wolfe (1970) of a mixture of multivariate normals, it is conjectured that the upper percentiles of a chi square distribution, with the number of degrees of freedom suggested by Wilks (1938), may provide an acceptable approximation to the large-sample distribution of $-2 \ln (LR)$ for mixtures.

Additional data collection and analysis related to the SIDS example in §3 may identify etiologic clues to aid the understanding of the sudden infant death phenomenon. If high rates were observed in counties of Virginia and South Carolina adjacent to the boundaries of North Carolina where regionalization occurred, then an environmental SIDS risk factor could be pursued. On the other hand, the absence of an elevated SIDS rate in these adjacent counties suggests possible political elements, perhaps related to differences in the delivery of health care or the reporting of SIDs. Supplemental data collection for the counties of North Carolina, seeking geographic, socio–economic and demographic correlates of the high SIDS rates, is nearly complete, and additional analyses are planned.

There are other potential applications for this methodology. Clustering of birth defects, cases of an infectious disease, or site-specific cancers may provide etiologic clues where there

have previously been none. For example, clustering of site-specific cancers may suggest an environmental hazard or occupational exposure.

Generalization of this approach to encompass the investigation for time–space clustering is straightforward, but computationally burdensome. The noninteractive time–space model would be a mixture model (2) with rates λ_g required to be constant across time periods for each county. A fully interactive time–space model would permit the rates to vary differently by time period for each county. A likelihood ratio test could be used to assess statistical significance. A restricted interactive model that allowed, for example, separate log-linear trends in the rates over time for each county would reduce the computations involved in fitting such a complex model, but might still provide a practical model for one form of time–space interaction.

Acknowledgements

The questions and helpful suggestions of two referees and the Associate Editor resulted in several improvements in this paper. The cooperation of the North Carolina Department of Human Resources is also gratefully acknowledged. Suspicion that there may be geographic clustering of the counties in North Carolina with high SIDS rates was brought to the attention of M. J. S. by Dr D. B. Gillings. Some independent work had already been completed by R. C. G. The research reported here was supported by an Intergovernmental Personnel Act appointment while M. J. S. was on leave with the Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina.

Résumé

Le regroupement d'observations d'une maladie rare est étudiée. On suppose que le nombre d'évènements observés pour chaque unité a une distribution de Poisson, dont la moyenne dépend de la taille de la population et de l'appartenance au groupe de cette unité. Ici un groupe est représenté par les unités qui sont homogènes pour leur taux d'occurrence des évènements rares étudiés. On modélise un échantillon d'unité par un mélange de distributions de Poisson, une pour chaque groupe, les paramètres du mélange sont les proportions de toutes les unités représentées par les composantes de la mixture. Les approches du maximum de vraisemblance et de Bayes sont utilisées pour déterminer les critères de séparation d'un échantillon en groupes d'unités à taux homogène. On présente en exemple, un test du rapport de vraisemblance pour la signification d'un mélange à deux composantes. On illustre la performance du critère avec des données d'occurrence spatiale de morts soudaines d'enfant de comtés de Caroline du Nord sur une période de quatre ans. Les résultats suggèrent qu'il serait bon de remettre en question la pratique courante qui consiste à diviser les comtés en catégories à haut ou bas risque sur la seule base des taux classés par ordre décroissant. Des tests basés sur des méthodes combinatoires sont aussi présentés pour examiner la signification du nombre de comtés contigüs parmi ceux à taux élevé de mort soudaine.

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

Atkinson, D. (1978). Epidemiology of sudden infant death in North Carolina: Do cases tend to cluster? PHSB Studies, No. 16. Raleigh, North Carolina: N.C. Department of Human Resources, Division of Health Services, Public Health Statistics Branch.

Batty, M. and Sammons, R. (1978). On searching for the most informative spatial pattern. *Environment and Planning, Series A* **10**, 747–779.

Binder, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

Cohen, A. C. (1963). Estimation in mixture of discrete distributions. In *Proceedings of the International Symposium on Discrete Distributions*, 373–378. Montreal.

Cox, D. R. and Hinkley, D. V. (1975). *Theoretical Statistics.* London: Chapman and Hall.

Ederer, F., Myers, M. and Mantel, N. (1964). A statistical problem in space and time: Do leukemia cases come in clusters? *Biometrics* **20**, 626–638.

Everitt, B. S. (1979). Unsolved problems in cluster analysis. *Biometrics* **35**, 169–181.

Geisser, S. (1966). Predictive discrimination. In *Multivariate Analysis*, P. R. Krishnaiah (ed.), 149–163. New York and London: Academic Press.

Goldberg, J. and Stein, R. (1978). Seasonal variations in SIDS. *Lancet* **2**, 107.
Grimson, R. C. (1979). The clustering of disease. *Mathematical Biosciences* **46**, 257–278.
Grimson, R. C., Wang, K. C. and Johnson, W. C. (1981). Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome. *Social Science and Medicine* **15**, 287–293.
Haggett, P., Cliff, A. D. and Frey, A. (1977). *Locational Models (Vol. 1)* and *Locational Methods (Vol. 2)*, 2nd eds. New York: Wiley.
Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon.
Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Boston: Houghton Mifflin.
Klauber, M. R. (1975). Space–time clustering analysis. A prospectus. In *Epidemiology*, D. Ludwig and K. L. Cooke (eds.), 75–89. Philadelphia: SIAM.
Knox, G. (1963). Detection of low intensity epidemicity—application to cleft lip and palate. *British Journal of Preventive and Social Medicine* **17**, 121–127.
Knox, G. (1964). The detection of space–time interactions. *Applied Statistics* **13**, 25–30.
Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference.*, Cambridge: Cambridge University Press.
Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.
Mantel, N., Kryscio, R. J. and Myers, M. H. (1976). Tables and formulas for extended use of the Ederer–Myers–Mantel disease-clustering procedure. *American Journal of Epidemiology* **104**, 576–584.
Marriott, F. H. C. (1975). Separating mixtures of normal distributions. *Biometrics* **31**, 767–769.
McRae, D. J. (1971). MIKCA: a Fortran IV iterative K-means cluster analysis program. *Behavioral Science* **16**, 423–424.
Ohno, Y., Aoki, K. and Aoki, N. (1979). A test of significance for geographic clusters of disease. *International Journal of Epidemiology* **8**, 273–281.
Rider, P. R. (1961). Estimating the parameters of mixed Poisson, binomial, and Weibull distributions by the method of moments. *Bulletin of the International Statistical Institute* **38**, 1–8.
Snow, R. H. (1968). Algorithm 342: Generator of random numbers satisfying the Poisson distribution. *Communications of the Association for Computing Machinery* **11**, 819.
Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* **37**, 35–44.
Wallenstein, S. (1980). A test for detection of clustering over time. *American Journal of Epidemiology* **11**, 367–372.
Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* **9**, 60–62.
Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* **5**, 329–350.

Received October 1980; *revised June and November* 1981

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

This article documents research that was conducted under the auspices of the U.S. Environmental Protection Agency (USEPA). As an arm of the government, the USEPA has a paid up, non-exclusive, irrevocable, worldwide license to reproduce, prepare derivative works, and distribute to the public all of the material contained herein, and to grant others acting on its behalf the right to do so. The USEPA grants to the National Technical Information Service, Department of Commerce, the right to act as its agent in the distribution of this material.